edWisor

Basic Statistics

- MUQUAYYAR AHMED
DATA SCIENTIST

Agenda We Learnt...

Statistics basics

- Introduction to Statistics
- Different areas of statistics
- Statistics Jargon

Today's Agenda

Central Tendency

Central Tendency

What is a measure of central tendency?

A measure of central tendency is a descriptive statistic that describes the average, or typical value of a set of scores

- Three Common Measures of Central Tendency
  - Mean
  - Median
  - Mode

Mean

The mean is:

- the arithmetic average of all the scores $(\Sigma X)/N$

The mean of a population is represented by the Greek letter **μ**; the mean of a sample is represented by X

You should use the mean when

- the data are interval or ratio scaled
- the data are not skewed

MEDIAN

The median is simply another name for the 50th percentile

The median is often used when the distribution of scores is either positively or negatively skewed

It is the score in the middle; half of the scores are larger than the median and half of the scores are smaller than the median

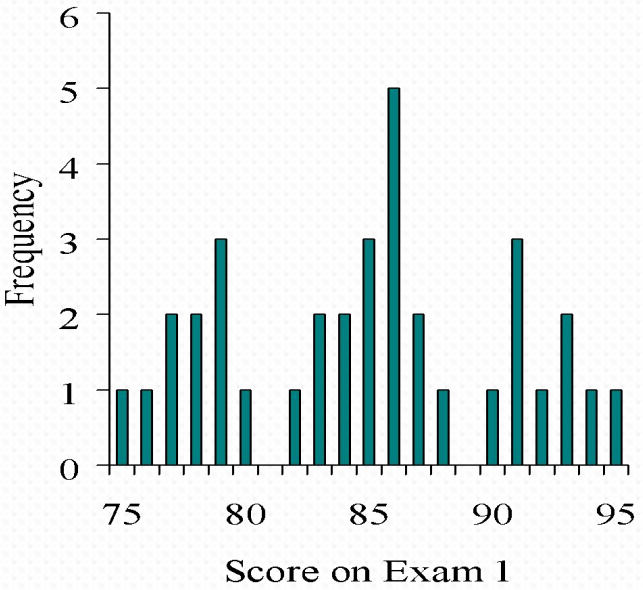
The few really large scores (positively skewed) or really small scores (negatively skewed) will not overly influence the median

Mode

The mode is the score that occurs most frequently in a set of data

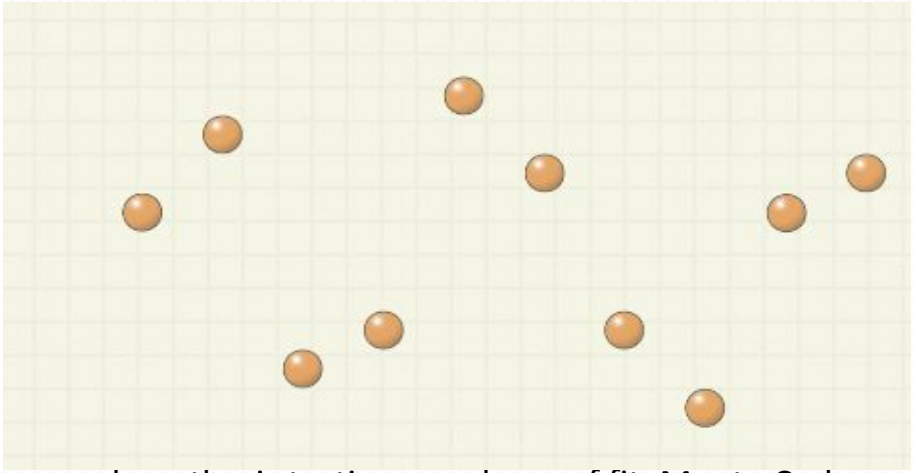
The mode is primarily used with nominally scaled data

- It is the only measure of central tendency that is appropriate for nominally scaled data



Variance

- Variance measures how far a set of numbers are spread out from their mean



- Variance can be calculated as

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

- It is used in descriptive statistics, statistical inference, hypothesis testing, goodness of fit, Monte Carlo sampling, amongst many others.

Standard Deviation

- Standard deviation tell you how much data deviates from the actual mean
- It is the square root of the Variance

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

- A low standard deviation indicates that the data points tend to be close to the mean, while a high standard deviation indicates that the data points are spread out over a wider range of values.